Out of the Scratch, Into the Blue? Considerations upon Designing of a New Digital Library

Henryk HOLLENDER
Warsaw University of Technology, Warsaw, Poland

INFORUM 2006: 12th Conference on Professional Information Resources
Prague, May 23-25, 2006

Abstract. *A new digital library, when carefully designed for an organization/community and compiled of the top quality materials, can almost replace a traditional or hybrid library, usually burdened with amassed useless documents. Few digital libraries, however, are transparent and articulate on their expected target audience and the subject structure of their contents. The strategic dimension of those libraries is often overwhelmed by the "scan/license what you can" approach, the budgeting restrictions blur vision and discourage strategic planning, and the flexibility of tools employed makes the editorial team endlessly postpone important decisions on functionalities and appearance of the database, thus inviting bad searching habits. Numerous small digital libraries around may for long remain heavily underused and misunderstood, validating the ironic "publish and perish" slogan.*

The Digital Library developed within Poland's biggest university of technology (Warsaw) is a case study of how to avoid this fate. The stress is put on original limitations of the project, the navigation techniques and metadata set adopted, and the prospects for the growth against the background of Poland's patterns in library policy and library cooperation.

By establishing a new digital library we seldom get engaged into a real planning, or at least, the planning as recommended by *A Framework of Guidance for Building Good Digital Collections* (NISO).[1] The miraculous technology available encourages us to produce sample lists in various formats, trying several arrangements and displays. These exercises, reaching the stage of recording and/or Internet transmission can almost imperceptibly take shape of a service. Testing software possibilities and limitations is a source of pleasure. It has already been paid for, and this is the machine which takes the drudgery over. The most arduous work of analogue librarianship — input of metadata — seems inessential when the target file can be easily located, and as a rule it can, because its very digital existence provides at least some access points.

The Internet is full of such immature services. Some undergo go the complex process of refinement and conversion, growing to greatness — or at least decency. Those which refuse to grow tend to stay over anyway in their substandard version. While sometimes marked "under construction", they are never marked "the worst practice", and they keep disseminating bad habits in information retrieval. They have one good thing about them: they do display some texts and thus can have their grateful users, as S. R. Ranganathan's third law of library science ("Every book has its reader") tells us. Actually and strongly, they operate in accordance with all the five of them.

The history of the Collection of Local Electronic Holdings (CLEH) launched in 2005 at the Warsaw University of Technology reflects some of these dilemmas. Now, available as Warsaw University of Technology Digital Library (http://bcpw.bg.pw.edu.pl), in the days of INFORUM 2006 it no doubt remains the "under construction" product. Even after all the tests, adjustmenst and amendments, it may long remain the newest and the smallest digital library of all ten supported by dLibra Digital Framework software supplied by Poznan Supercomputing and Networking Center. The project is important because the Warsaw University of Technology is the largest technical school in Poland, and because Warsaw is

the capital which until now has not had its regional digital library — in contrast to provinces like Cracow or Wrocław. Two most extensive digital projects in Poland do have headquarters in Warsaw, namely Polska Biblioteka Internetowa (Polish Internet Library, PBI, Ministry of Science) and Biblioteka Wirtualna Nauki (Virtual Library of Science, BWN, Warsaw University), but other digital initiatives in the country's major center of tertiary education are not digital libraries, because they just services unsupported by a specialized software. They generally belong to the field of library automation, like NUKAT — the national union catalogue project. When PBI, dedicated to digitization, failed to secure basic quality standards, and BWN, dedicated to just everything, with a stress on international scientific contents, failed to show its long-promised search engine, save encompass any major contents produced by digitization, there was a feeling that Warsaw was underrepresented on the digital scene of Poland. At the same time dLibra libraries tend to grow, and some day an obvious urge may cause the Warsaw University of Technology DL to turn to a Digital Library of Mazovia. And by Mazovia we mean a province with over 19% of the country's student population of almost 2 million.[2]

Thus the WUTDL could have been designed as a large library. But it was not. It was designed in the late 2004 as a small library, causing not much pressure on the Warsaw University's of Technology budget. It was obvious that it may grow some day. For the director of the library, experienced in library automation, it was probably clear that in the future most of articles, books, music, images and motion pictures would be located and read online. There was no discussion that there is only one direction in the world of communications. The fate of works of scholarship which will have somehow avoided the digital format will be similar to those manuscripts which once avoided going to press. But the Main Library of a major technical university is not a single agent of change, and the director might have shared with her colleagues the intuition that we in Poland need not establish and augment regional digital libraries if we still have a chance for and faith in the BWN as the big and universal one. Thus the roots of this digital library are modest: originally it started as a HTML list of works from the Library's own printed format collections, digitized for

1. better access, and
2. increased awareness of the school's history.

Some overused handbooks, many published twenty or thirty years ago and still assigned students as obligatory, were selected to meet the first requirement. The second one was ideally fulfilled by typewritten (or handwritten!) lecture notes mimeographed by students union (Towarzystwo Bratniej Pomocy Studentów) in the 1920s, quickly supplemented by other important works from the early decades of the school, established in 1898. Most of the work was done on a Minolta PS 7000 scanner, with book covers copied on a colour office scanner. Results were displayed in PDF and HTML format. The service was still available on the Internet prior the conference (http://www.bg.pw.edu.pl/elib).

The project was launched in early 2005, soon after antiquarian collections of the Library were identified according to newly adopted rules on the "national heritage". It was clear that some of those items will not only be cleaned, repaired, bound etc., but also digitized — all that for better protection. Experiments were then started with digital republishing of those items. Eventually, the CLEH consisted of three collections:

- History of Science and Technology
- History of Warsaw University of Technology
- Texts and Handbooks,

and rough list of items for inclusion in each was compiled. Now, the DL is to have more collections — in addition to abovementioned:

- Illustrated History of Technology
- Illustrated History Architecture and Construction
- Scientific Repositorium
- Serials (complete runs).

These have emerged recently, with acquisitions of a rich pictorial collection inherited after one liquidated library in Warsaw. The repositorium has already its first items under preparation, but its existence and growth depends on the schools decision of adopting an open access policy on master's theses, PhD dissertations and postdoctoral research. A map collection is also envisaged, and some of the collections may in the future have not much to do with the Warsaw University of Technology's otherwise far-flung programmes. Then it perhaps really does turn into the digital library of the province of Mazovia — but this not without some pity that one central library for the country has eventually not come into being. To have this achieved, however, we would have to have an institutional center, which we apparently and visibly are lacking.

It may be then said that the collection of Local Electronic Holdings, now the Digital Library, was originally established to:

- open up a new prospect and keep the possibilities open
- feed the need of the school for better access to teaching materials
- celebrate the school's past and invite a research into its history.

It would also be beautiful it the service showed the charm of the history of science and technology — a discipline gone almost for good from Polish universities. But the project was imbued with self-limitation. During the first year of the project there was no full-time staff for selection, scanning, and processing. Some variant of Dublin Core was adopted for metadata, but it only visually arranged the data as the RDF format was not introduced. Most parameters were chosen intuitively to secure the acceptable performance. The slow growth of the collection in a sense permitted the prolonged operation of the library without search and retrieval tools. Browsing of the lists for each collection or of the union list of the items had (and did) suffice.

What was paid much attention to was the legal aspects of the project. We announced that we would publish or republish digital-born works which authors used for the courses they thought, and the effort was made to notify all the faculties. Only two works were submitted, both authored by an instructor who tought in English. We discussed the lists of heavily used texts with two major publishing houses: Oficyna Wydawnicza Politechniki Warszawskiej (Warsaw University of Technology) and Wydawnictwa Naukowo-Techniczne (Technical and Scientific Publishers). Both informed us of works they were no longer interested, and accepted our turning directly to the authors or their heirs. These were suggested to sign with us a licence for digitization and republishing of the work for no fee, and with a right to request us to withdraw it if a chance for a "regular" publishing emerged. Thus we came to own right for several texts and handbooks, whose authors were often surprised that the development of new technologies and new scientific findings had not render their works useless. We cannot but believe that we keep offering classics, a hard core of knowledge which is not yet to be abandoned, but we would love to be as successful as Akademicka Biblioteka Cyfrowa in Cracow (http://abc.agh.edu.pl) or Dolnośląska Biblioteka Cyfrowa in Wrocław

(http://www.dbc.wroc.pl/dlibra) in acquiring new publications of the local academic press or just the digital born works by the faculty.

In publishing of the well-tested works on the Internet, we only exceptionally let authors amend their texts. Instead, we invite authors to write a short introduction which is displayed next to the bibliographic description of the work, and which informs the user to what extent the modern knowledge differs to that represented in the work. We also electronically retouch the cover, which normally is pretty worn out. As for antiquarian items, we clean up some of the original's dirt, but also correct mistakes of the binder. We also use a Casio QV5700 digital camera for items in which colour should be shown. With dLibra software, we are going to display the images of covers of nearly all the books for the users' orientation. If the binding of an antiquarian items does not provide any information, be it edition-specific or copy-specific, we will offer a colour image of the title page, which usually shows stamps and annotations. With modern items held in the library in their library bindings we will "create" new bindings electronically form the title page. We can and will supplement missing pages by including digital material derived from various copies. Thus we think we create a new document, which is not identical to the printed original. This affects our thinking of cataloguing.

To explain this, we have to return to the beginnings of digitization in Poland. Since small-scale projects dominated, and there was no body to coordinate and control scattered initiatives, EBIB service [http://www.ebib.info] maintained and updated a database of projects launched (2002-2003). Despite the progress, few new projects were submitted to the database. It ceased to play any role and disappeared. Then many put their faith in the emerging NUKAT national union catalogue project, pointing to it that contributing to NUKAT of a record for any digitized item would secure the full control of the progress of digitization and help avoid duplication. But many libraries ignored NUKAT, and the issue remained open.

As for the Main Library of the Warsaw University of Technology, it started entering its MARC 21 records for electronic documents to NUKAT. As a next step, those records were downloaded to the school's online catalogue. A user could then easily locate the CLEH documents via Google, via NUKAT, and via the local online catalogue (Aleph 500), with its numerous access points. The CLEH offered a link to the record, both from the level of the HTML list which served as the title index to the documents collected, and from the level of the full bibliographic (quasi-DC) record. It was very rewarding to notice that the regular online catalogue of the Nicolaus Copernicus University (Horizon) offered link to items in CLEH — they could also be seen in NUKAT as one of the "locations" of the electronic document. Eventually (and following the British Library Website Copyright Statement)[3] we announced that we welcome linking. Moreover, our records for electronic documents in the online catalogue were mutually linked to the records for original documents, and, occasionally, to CD ROM versions which were made at the Library before the decision on Internet publishing was taken. On the top of that, we started linking from our online catalogue to useful publications we located in the other digital libraries, whether we had the printed version or not.

We gather the opinion that the users ignore online catalogues when given a chance of searching or browsing of a digital library, and that in the future they will tend to confine themselves to digital libraries for good. We still don't know, however, to what extent this attitude is caused by the lack of full-text linking in normal online catalogues. When digital libraries fully support learning and have sophisticated search capabilities, we can dissociate

them from online catalogues and leave the latter for the use of scholars — like we did with card catalogues when online catalogues took root. For the time being, however, we think that hybrid access to electronic documents, when only clearly explained to the users, will augment the usage and fairly reflect the dual character of our holdings.

Thus we, in a sense, accepted the slow growth of our digital collection. Achieving of a usable amount of documents which would permit us giving it a name of "library" would take a special project with a generous grant. You can quicken up the selection and procession of items for digitization, but you are not likely to return to records to amend them.

This was also the reason why we carefully selected the Dublin Core fields for records within the new Digital Library. The most sensitive issue was the year of publication. All dLibra libraries thought that this is the original year of publication what the users is interested in, along with the original publisher, because the electronic format was accidental and external to the contents and appearance they sought for. This thinking went along the line: we give you a digital copy instead of, say, an early imprint. Therefore in most of those records the year of the digital republication does not show at all. This author thinks, however, that when accessing a digital library we deal with electronic documents, and not printed books, manuscripts or photographs, and that the electronic format is not external and neutral to the contents.[4] This applies not only to the material which just does not have its paper equivalent, which is what we want to publish in large quantities, but also to any republication which brings about edition, supplement, or transformation of a the paper original (or, to be precise, of a digital copy of the paper original). No, we are not just copying of the printed (or manuscript, or pictorial) heritage! Some of our digital documents are compilations of several imperfect paper copies (or precisely, of several digital documents as the imperfect paper copies were scanned), the other carry remarks by authors, which also belong to what is being published. We want you to precisely trace the paper document if it is the source of what is now being viewed, but we also want you to expect the added value from the digital version. If your focus is on the digital version, this actually is what should be catalogued. You do it by describing it with a clearly defined metadata set for the digital document.

Thus we came to argue that the year of publication is "now", i. e. when the electronic document goes online, and we keep having at our disposal the Source field which lets us furnish the user with all the information about the item published "in the beginning". Eventually, to compromise with the practice other dLibra libraries, we decided that the year of publication is nevertheless "then" (instead of "now" or "recent"), but we call our Date field "Original publication date" (Data wydania oryg.)". Under Source the user receives the link to the full MARC record for the original, and under Rights — the year of the digital publication. All in all, we believe we developed a simple and informative dataset.

The other element absolutely necessary for the proper orientation of the user is the text navigation. In the future the contents of this Digital Library will be fully covered by full text searches. It takes implementation of DjVu format and OCR technology. For the time being we felt that, while continuing using PDF and HTML (mostly PDF) format, we cannot deprive the user of the right to select the part of the publication in which an expected subject matter was dealt with. The Warsaw University of Technology keywords may help locate a document, but what next? The mainstream practice is just unacceptable. It forces the user to view the document page by page, or to jump to a specific page. If you select a page number from the table of contents which is a part of the digital document, and this is the number you select, you probably are mistaken as the system counts subsequent images and not actual original

pages. If the original document has five unnumbered pages at the beginning, you may want the page number 15, but in fact you get the "page" number 10, and so on. It is strange that the issue receives so little attention. We can expect that soon automatic orientation, navigation, indexation and retrieval tools will be developed, and this may be a crop of the eContent*plus* program. For the time being, however, we deal with thousands of electronic books which are extremely difficult to use, and which seem to support only some kinds of superficial examination, and not really study or research.

You can buy a CD ROM which is neatly indexed (like CD Retieval indexation employed in CD ROMs published and sold by the Warsaw University Library[5]), but we do not normally see the equivalents on the Internet. The provisional response of the Warsaw University of Technology was dividing the document into convenient files (normally up to 20 pages) identical to chapters of a book. The chapters could be seen on the table of contents before the document was viewed. The original chapters could be employed with a hypertext quality, so it was enough to click on the chapter title or subtitle to open it. Otherwise, a special table of contents was developed. This was usually necessary with antiquarian materials, which had very obscure table of contents, illogical and incompatible with the body of the text. We felt that there was a tension of a kind between the original table of contents, the actual text structure, and the running pages, and it could be researched by a book historian. Designing of an "artificial" and "clickable" table of contents, which you can see on our pages, was an arduous task, but it reflected some of the lure of such a research as it required analysis of how the narration in the book was structured.

Now, supported with dLibra software, we are one more "typical" digital library of Poland. They soon will be able to be all searched in one session, which will bring about a very substantial improvement of access to the written heritage of the country. They offer all the regular browse and search functions — basic, advanced and Boolean, and they are heading towards full-text searches. They need constant care and improvement, and this will not happen easily without more strict cooperation in the framework of a consortium. Majority of their collections is digitized works, which will mostly serve as historical evidence, to a limited number of scholars. Adding to the collections a strong body of contemporary science, peer-reviewed research findings, reference works and study aids for students will decide their future and the size of their audience.

And the audience remains mostly silent. The feedback for the emerging Warsaw University of Technology Digital Library was very small. Instructors did not rush in with their works. Students Union informed us of the works needed badly for study, thus reinforcing our own lists and statistics, but no grateful readers sent in their gratitude when the works were included. No user made a comment of our navigation apparatus that we took so much effort to offer. The local e-learning center was not helpful. The reserve room never existed, and there is no common will to have one established as an additional service of the Main Library. Letters, posters, flyers and library orientation classes in the new academic year may draw a number of aware users, and well may not. In the big schools, the mainstream support for studying remains just circulating thousands of printed titles in the attractive milieu, which includes extended modern library premises. Looking for stakeholders has only begun, and is far behind the technical progress of digitization.

The digital library will not fell into oblivion, but in a short run it may not seem a worthwhile investment, and in the long run it may suffer from a competition of commercial digital libraries, which will certainly show up and become affordable for many students, if not for the

school. Thus some danger and uncertainty exists. A very careful selection of items for digital publication/republication, combined with even more friendly interface is the only imaginable answer. If we fail to organize a service very heavily used, heavily sought for, it must be the blue sky and not the hard reality we looked at. No matter how careful and pleasurable the designing has been.

---

[1] Cf. M. Nahotko: *Zasady tworzenia bibliotek cyfrowych* [Rules for designing of digital libraries],[online document], "Biuletyn EBIB" 2006 nr. 6 (April), http://www.ebib.info/2006/74/nahotko.php. Accessed on May 15th , 2006.

[2] *Szkoły wyższe i ich finanse w 2004 r.* [Tertiary education and how it is budgeted],[online document], in: Główny Urząd Statystyczny, http://www.stat.gov.pl/dane_spol-gosp/warunki_zycia/szkoly_wyzsze_w_2004/index.htm, table 3: Studenci szkół wyższych według województw. Accessed on May 15th , 2006.

[3] http://www.bl.uk/copyrightstatement.html. Accessed on May 15th , 2006.

[4] Cf. http://dublincore.org/documents/usageguide, 1.2.1. Accessed on January 24th , 2006

[5] Cf. the list in the Shop information, http://buwcd.buw.uw.edu.pl/sklep/sklep.htm. Accessed on May 4th, 2006.